FILTERING INPUT DATA ERRORS IN MODEL ERROR CALCULATIONS: IMPACT OF MEASUREMENT UNCERTAINTY ON MEAN ABSOLUTE ERROR ESTIMATION

DRAFT VERSION

Vasilii Piiadov

piyadov@alumni.usp.br

Last update: July 24, 2025

ABSTRACT

This paper examines the impact of measurement uncertainty on Mean Absolute Error (MAE) calculations in statistical analysis and model validation. We derive analytical expressions for the distribution of absolute differences between observations under various uncertainty models, including delta function distributions for data without uncertainty and normal distributions for measurement errors. Our analysis demonstrates that conventional MAE calculations can lead to systematic overestimation when input data contains uncertainty, particularly in cases where the signal-to-noise ratio is low. We establish sufficient conditions under which standard MAE calculations remain approximately valid and propose methods for accurate MAE estimation in the presence of measurement uncertainty. The findings have significant implications for model validation in industrial applications where measurement precision affects performance assessment.

1 Introduction

Consider two sets of unknown true values $\{a_k\}$ and $\{b_k\}$ that are estimated through observations $\{x_k\}$ and $\{y_k\}$, respectively. The Mean Absolute Error (MAE) between the true values is defined as:

$$h = \frac{1}{n}\sum_{k=1}^{n} |a_k - b_k| \tag{1}$$

In practical applications, we seek to calculate the MAE using the available dataset of observations $\{x_k\}$ and $\{y_k\}$. However, the relationship between observed and true MAE depends critically on the uncertainty characteristics of the measurement process.

2 Distribution of absolute difference between observations

Let $x \in X$ and $y \in Y$ represent single estimations of a and b, respectively, modeled as random variables. We denote $F_X(x)$ and $F_Y(y)$ as the distribution functions describing the uncertainty in single observations, with corresponding probability density functions $f_X(x)$ and $f_Y(y)$. Note that these distribution functions depend on the true values a and b such that x and y serve as their estimators.

We consider the absolute difference z between x and y: $z \in Z$, where $Z = |X - Y|$ and $z \geq 0$.

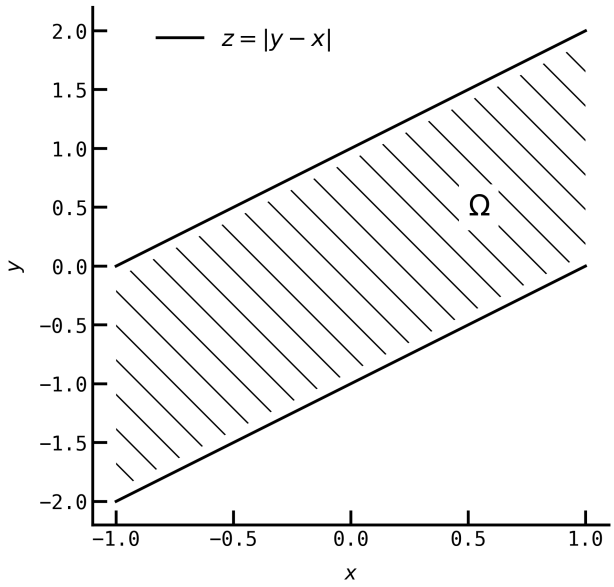

Figure 1: Joint density function domain Ω

Let $F_Z(z)$ denote the distribution of the absolute difference in a single observation. The distribution function is given by:

$$F(z) = \iint_{\Omega} f_X(x) f_Y(y)\, dy\, dx = \int_{-\infty}^{\infty} \int_{x-z}^{x+z} f_X(x) f_Y(y)\, dy\, dx,\ z \geq 0 \tag{2}$$

$$f(z) = \frac{dF(z)}{dz} = \int_{-\infty}^{\infty} f_X(x) \left(f_Y(x+z) + f_Y(x-z)\right)\, dx,\ z \geq 0 \tag{3}$$

In the following sections, we examine important special cases of uncertainty distributions and their relationship to the absolute difference $|a-b|$ and the MAE h from Equation 1.

3 Data without uncertainty

We first consider the idealized case of data without measurement uncertainty. In this scenario, the probability density functions for observations are characterized by Dirac delta functions:

$$\begin{aligned} f_X(x) &= \delta(x-a) \\ f_Y(y) &= \delta(y-b) \end{aligned} \tag{4}$$

If $m = a - b$,

$$\begin{aligned} f_X(x) &= \delta(x) \\ f_Y(y) &= \delta(y-m) \end{aligned} \tag{5}$$

$$f(z) = \delta(z-m) + \delta(z+m),\ z \geq 0, \forall m \in \mathbb{R} \tag{6}$$

Characteristic function:

$$\psi(t) = \int_{0}^{\infty} e^{i\,t\,z}\, f(z)\, dz = e^{i\,|m|\,t} \tag{7}$$

Expected value and variance: *DRAFT VERSION*

$$\begin{aligned} E[Z] &= \frac{1}{i}\frac{d}{dt}\ln\psi\Big|_{t=0} = |m| \\ V[Z] &= -\frac{d^2}{dt^2}\ln\psi\Big|_{t=0} = 0 \end{aligned} \tag{8}$$

Equation 8 is consistent with the problem formulation: in the observation of random variables X and Y, we can only obtain $x = a$ and $y = b$. Therefore, Equation 6 can be interpreted as the uncertainty distribution for data without measurement uncertainty, and the proposition in Equation 4 is valid for this case.

The probability that the observed absolute difference z is less than some value z' is:

$$P(z \le z') = \int_0^{z'} f(z)\,dz = \int_0^{z'} \left[\delta(z-m) + \delta(z+m)\right]\,dz = H(z'-m) + H(z'+m) - 1, \tag{9}$$

where $H(x)$ denotes the Heaviside step function.

From Equation 10, it follows that the probability can only assume three distinct values:

$$P(z \le z') = \begin{cases} 0 & \text{if } |m| = 0,\ z' = 0 \text{ or } |m| > z' \\ \frac{1}{2} & \text{if } |m| = z',\ z' > 0 \\ 1 & \text{if } |m| < z' \end{cases} \tag{10}$$

Since we have observed z, the probability $P(z \le z')$ can only take values of $\frac{1}{2}$ or 1 (indicated by the shaded area in Figure 2).

The case $P(z \le z') = \frac{1}{2}$ represents a marginal uncertainty scenario because at $z' = |m| + \delta z$ we have $P(z \le z') = 1$, while at $z' = |m| - \delta z$ we have $P(z \le z') = 0$, where δz is an infinitesimal variation of z.

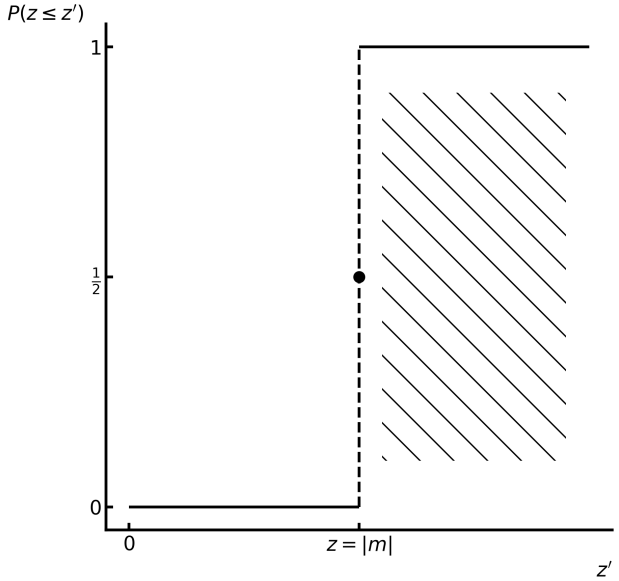

Figure 2: Probability of $z \le z'$. Patched area shows possible cases of z' for observed value z.

Since $z = |m|$ (Equation 8), and defining $\varepsilon = z' - |m|$ as the tolerance for comparing z and $|m|$ (equivalently, $|x-y|$ and $|a-b|$), the probability $P(z \le z')$ represents the likelihood that the values $|x-y|$ and $|a-b|$ can be considered equivalent within tolerance ε.

For a series of observations $\{x_k\}$ and $\{y_k\}$, we examine the integral characteristic expressed by the random variable:

$$W = \frac{1}{n}\sum_{k=1}^{n} Z_k \tag{11}$$

Using Equation 8 and the properties of expected value and variance, we obtain:

$$\begin{aligned} E\left[W\right] &= \frac{1}{n}\sum_{k=1}^{n} E[Z_k] = \frac{1}{n}\sum_{k=1}^{n} |m_k| = h \\ V\left[W\right] &= \frac{1}{n^2}\sum_{k=1}^{n} V[Z_k] = 0 \end{aligned} \tag{12}$$

The condition required by Kolmogorov's strong law of large numbers, $\sum_{k=1}^{\infty} \frac{V[Z_k]}{k^2} < \infty$, is clearly satisfied. Since $E\left[W\right]$ equals the MAE h (Equation 1), we have:

$$h = \frac{1}{n}\sum_{k=1}^{n} z_k = \frac{1}{n}\sum_{k=1}^{n} |x_k - y_k|, \text{ as } n \to \infty \tag{13}$$

Equation 13 represents the standard method for calculating MAE in practice. For the simple case of data without uncertainty (Equation 6), this approach coincides with Equation 1. While the same result could be obtained directly from the fact that $a_k = x_k$ and $b_k = y_k$ for data without uncertainty, we have derived it using a general approach that relies only on the uncertainty distribution (Equation 6) without explicit use of prior knowledge about the values of a_k and b_k. This methodology can therefore be extended to more complex cases involving uncertainty described by other distributions.

It is important to note that Equation 13 is valid only for data without uncertainty; for other uncertainty distributions, the relationship will not have such a simple form.

However, in many practical applications involving dataset comparisons using absolute differences (e.g., MAE for model performance evaluation), MAE is calculated using the form of Equation 13 even when data contains uncertainty. This approach is generally incorrect and can lead researchers to erroneous conclusions. In the following sections, we explore the impact of uncertainty for two cases of significant practical importance.

4 Data with normal distribution of uncertainty

In this case, we assume that the individual estimations x and y follow normal distributions (Equation 14) with means a and b, respectively. In practice, this scenario corresponds to most physical measurements obtained from sensors, measuring instruments, or mathematical/ML models where uncertainty in input data propagates to uncertainty in outputs. Thus, this case is applicable, for example, to the validation of such models against physical measurements in industrial projects (quality control, product property forecasting, etc.).

$$\begin{aligned} f_X(x) &= \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-a}{\sigma_x}\right)^2} \\ f_Y(y) &= \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-b}{\sigma_y}\right)^2} \end{aligned} \tag{14}$$

Using Equation 3, the uncertainty distribution $f(z)$ becomes:

$$f(z) = \frac{1}{\sigma\sqrt{2\pi}} \left[e^{-\frac{1}{2}\left(\frac{m+z}{\sigma}\right)^2} + e^{-\frac{1}{2}\left(\frac{m-z}{\sigma}\right)^2} \right] \tag{15}$$

where $m = a - b$ and $\sigma = \sqrt{\sigma_x^2 + \sigma_y^2}$.

By finding the characteristic function of the distribution in Equation 15 and computing the corresponding derivatives, we calculate the expected value and variance for this distribution:

$$\begin{aligned} E[Z] &= m \operatorname{erf}\left(\frac{m}{\sqrt{2}\,\sigma}\right) + \frac{\sqrt{2}\,\sigma\, e^{-\frac{1}{2}\left(\frac{m}{\sigma}\right)^2}}{\sqrt{\pi}} \\ V[Z] &= m^2 + \sigma^2 - E[Z]^2 \end{aligned} \tag{16}$$

As can be observed, Equation 16 is symmetric with respect to the sign of m, so m can be replaced by $|m|$. From the behavior of both components in the $E[Z]$ equation:

$$E[Z] \geq |m| \tag{17}$$

where equality $E[Z] = |m|$ is achieved only as $|m| \to \infty$. Note also from Equation 16 that $V[Z]$ is a monotonically increasing function, so $\max_{m \in \mathbb{R}} V[Z] = \lim_{|m| \to \infty} V[Z] = \sigma^2$.

Summing over the entire series $\{z_k\}$:

$$\frac{1}{n}\sum_{k=1}^{n} E[Z_k] \geq \frac{1}{n}\sum_{k=1}^{n} |m_k| \tag{18}$$

In the series $\{z_k\}$, maximal value of Kolmogorov's condition achieved at $\forall |m_k| \to \infty$ will be:

$$\lim_{\forall |m_k| \to \infty} \sum_{k=1}^{\infty} \frac{V[Z_k]}{k^2} = \sum_{k=1}^{\infty} \frac{\lim\limits_{|m_k| \to \infty} V[Z_k]}{k^2} = \sum_{k=1}^{\infty} \frac{\sigma}{k^2} = \frac{\pi^2\,\sigma}{6} < \infty \tag{19}$$

Easy to see, for any other $|m|$, sum from the left part of Eq.19 less than $\frac{\pi^2\,\sigma}{6}$. Thus, using Kolmogorov's theorem we have:

$$\frac{1}{n}\sum_{k=1}^{n} z_k \geq \frac{1}{n}\sum_{k=1}^{n} |m_k|, \text{ as } n \to \infty \tag{20}$$

From Equation 20, we observe that the mean of observed absolute differences is greater than the MAE and, in general, cannot be used as an MAE estimator, unlike the case of data without uncertainty considered previously. However, we can establish a sufficient condition for the approximation:

$$\frac{1}{n}\sum_{k=1}^{n} z_k \approx \frac{1}{n}\sum_{k=1}^{n} |m_k| \tag{21}$$

Due to the law of large numbers, Equation 21 can be written as:

$$\frac{1}{n}\sum_{k=1}^{n} E[Z_k] \approx \frac{1}{n}\sum_{k=1}^{n} |m_k| \tag{22}$$

We make the following change of variables:

$$u_k = \frac{|m_k|}{\sqrt{2}\sigma} \tag{23}$$

Therefore, the expression for $E[Z_k]$ (Equation 16) takes the form:

$$E[Z_k] = \sqrt{2}\sigma \left[u_k \mathrm{erf}(u_k) + \frac{1}{\sqrt{\pi}}\, e^{-u_k^2} \right] \tag{24}$$

Introducing a correction coefficient $\alpha \approx 1$, we can rewrite Equation 22 as:

$$\sum_{k=1}^{n} \left(u_k \mathrm{erf}(u_k) + \frac{1}{\sqrt{\pi}}\, e^{-u_k^2} \right) = \alpha \sum_{k=1}^{n} u_k \tag{25}$$

Differentiating Eq.25 with respect to each u_k:

$$\sum_{k=1}^{n} \mathrm{erf}(u_k) = n\,\alpha \tag{26}$$

For convenience we can use the same correction coefficient for each part of the sum from Eq.26. Thus,

$$\mathrm{erf}(u_k) = \alpha, \, \forall k \tag{27}$$

is sufficient for Equation 26 to be satisfied. Therefore, if $\alpha \gtrsim 0.95$, then $\alpha \approx 1$ in many cases.

Reverting the variable change (Equation 23), we obtain the sufficient condition for Equation 21:

$$|m_k| \gtrsim 2\sigma, \forall k \tag{28}$$

Therefore, Equation 21 can be applied only when the values in series $\{a_k\}$ and $\{b_k\}$ are sufficiently different (Equation 28). Depending on the context and the degree of uncertainty expressed by σ, this condition may limit the applicability of Equation 21. For example, in the context of model validation where $\{a_k\}$ represents model predictions compared against measurements $\{b_k\}$ under conditions of high data uncertainty (large σ), the applicability of Equation 21 will be limited to poorly performing models, which is insufficient for many applications.

Let us consider $E[Z]$ (Equation 16) as a function $\phi(m)$. Since $\phi(m)$ is a convex function, Jensen's inequality gives us:

$$\phi(h) \leq \bar{\phi}, \tag{29}$$

where $\bar{\phi}$ is the mean of $\phi(m)$ over $\{m_k\}$: $\bar{\phi} = \frac{1}{n}\sum_{k=1}^{n} z_k$ and $h = \frac{1}{n}\sum_{k=1}^{n} m_k$, which corresponds to the MAE definition in Equation 1.

We define ε as the relative difference between $\bar{\phi}$ and $\phi(h)$:

$$\varepsilon = \frac{\bar{\phi} - \phi(h)}{\phi(h)} \tag{30}$$

Using the maximum value of ε and Equation 29, we obtain:

$$\frac{1}{1+\varepsilon_{max}}\bar{\phi} \leq \phi(h) \leq \bar{\phi} \tag{31}$$

Therefore, the accuracy with which we can determine $\phi(h)$ from Equation 31, and hence the MAE $h = \phi^{-1}(\phi(h))$, depends on ε_{max}, which we can estimate through numerical experiments.

Consider values from $\{m_k\}$ constrained to an interval $[\mu_1, \mu_2]$ and distributed uniformly. By varying the values of $\mu_2 \geq \mu_1 \geq 0$, we can calculate possible values of ε and estimate its maximum value ε_{max}. The uniform distribution of $\forall m_k \in [\mu_1, \mu_2]$ is a good modeling choice for simulation because of its heavy tails; this provides a worst-case estimation of ε compared to distributions with lighter tails, since greater dispersion over $[\mu_1, \mu_2]$ enhances the inequality in Equation 29 over the interval.

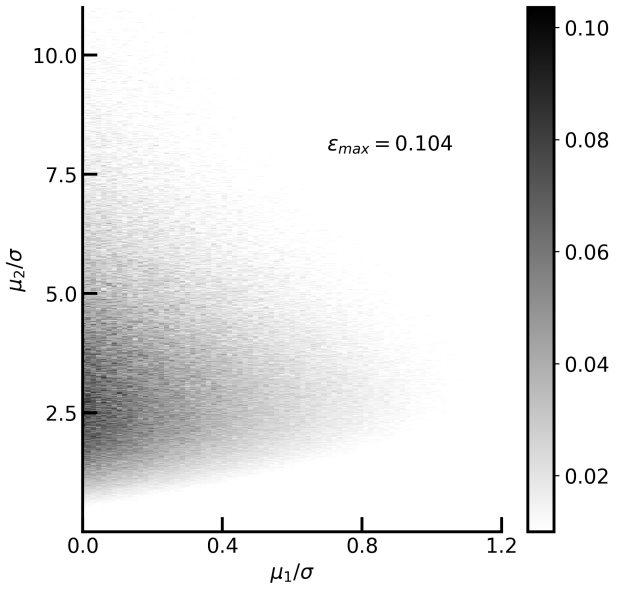

Figure 3

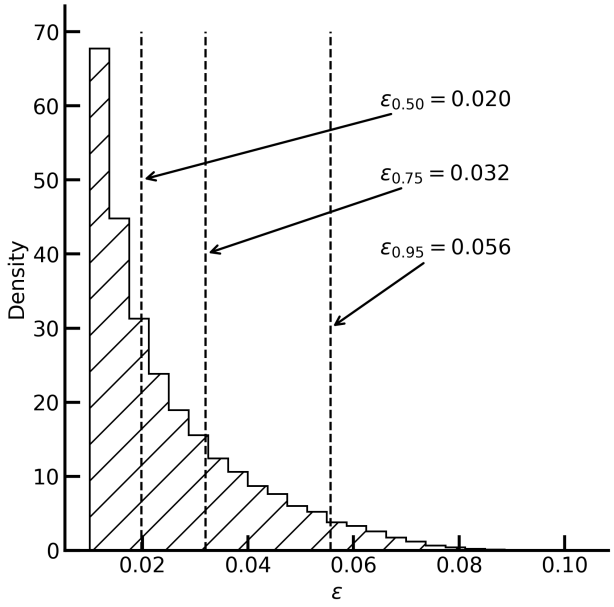

Figure 4

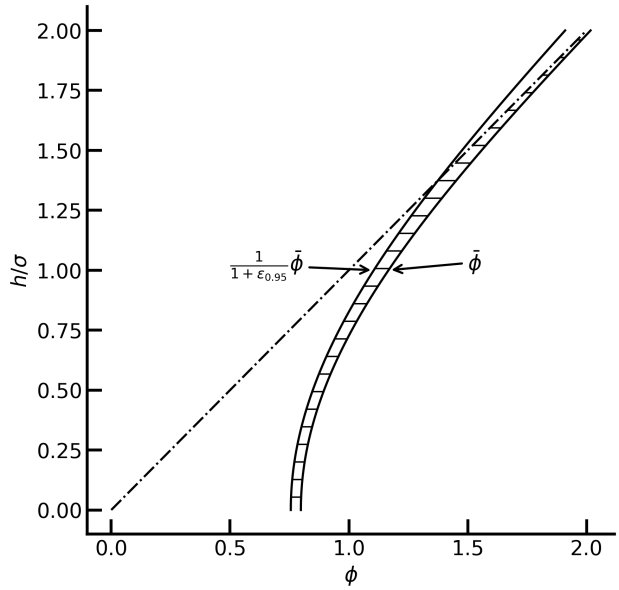

Figure 5: Accuracy of MAE h determination

5 Mixed case: Market price dispersion example

Consider the example of market price dispersion where different sellers have different offers for the same commodity. For instance, multiple real estate agents may want to sell the same house, each with their own selling strategy, resulting in prices distributed according to some probability distribution. In this scenario, prices are random variables.

Now imagine that a dataset contains prices from only one real estate agent. In this case, the true price, which should be determined by objective market circumstances, is unknown. Typically, predictive models are based on market indices rather than the psychology of a selected real estate agent, highlighting the importance of understanding measurement uncertainty in such applications.

7 Acknowledgments